

REPORT DOCUMENTATION PAGE			Form Approved OMB NO. 0704-0188		
<p>The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</p> <p>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</p>					
1. REPORT DATE (DD-MM-YYYY) 10-10-2017		2. REPORT TYPE Final Report		3. DATES COVERED (From - To) 1-Apr-2015 - 30-Sep-2015	
4. TITLE AND SUBTITLE Final Report: PREDICTOR - Predictive REaction Design via Informatics, Computation and Theories of Reactivity			5a. CONTRACT NUMBER W911NF-15-1-0041		
			5b. GRANT NUMBER		
			5c. PROGRAM ELEMENT NUMBER		
6. AUTHORS Dean Tantillo, Uttam Tambar, David Wild			5d. PROJECT NUMBER		
			5e. TASK NUMBER		
			5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAMES AND ADDRESSES University of California - Davis Sponsored Programs 1850 Research Park Drive, Suite 300 Davis, CA 95618 -6153			8. PERFORMING ORGANIZATION REPORT NUMBER		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES) U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211			10. SPONSOR/MONITOR'S ACRONYM(S) ARO		
			11. SPONSOR/MONITOR'S REPORT NUMBER(S) 66705-CH-DRP.1		
12. DISTRIBUTION AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.					
13. SUPPLEMENTARY NOTES The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.					
14. ABSTRACT The goal of this program was to create a cyber-expert software PREDICTOR capable of interacting with a synthetic organic chemist to expedite the discovery of new chemical reactions. Employing state of the art quantum chemical reactivity modeling methods combined with a scalable high-performance reaction database, we began to design an artificially intelligent software that will store experimental reactivity data and expand its knowledge base using computer simulations.					
15. SUBJECT TERMS computational chemistry, reaction design					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Dean Tantillo
a. REPORT UU	b. ABSTRACT UU	c. THIS PAGE UU			19b. TELEPHONE NUMBER 530-757-2528

RPPR Final Report

as of 24-Nov-2017

Agency Code:

Proposal Number: 66705CHDRP

Agreement Number: W911NF-15-1-0041

INVESTIGATOR(S):

Name: Dean J. Tantillo
Email: djtantillo@ucdavis.edu
Phone Number: 5307572528
Principal: Y

Organization: **University of California - Davis**

Address: Sponsored Programs, Davis, CA 956186153

Country: USA

DUNS Number: 047120084

EIN: 946036494

Report Date: 31-Dec-2015

Date Received: 10-Oct-2017

Final Report for Period Beginning 01-Apr-2015 and Ending 30-Sep-2015

Title: PREDICTOR - Predictive REaction Design via Informatics, Computation and Theories of Reactivity

Begin Performance Period: 01-Apr-2015

End Performance Period: 30-Sep-2015

Report Term: 0-Other

Submitted By: Dean Tantillo

Email: djtantillo@ucdavis.edu

Phone: (530) 757-2528

Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees:

STEM Participants:

Major Goals: The goal of this program was to create a cyber-expert software PREDICTOR capable of interacting with a synthetic organic chemist to expedite the discovery of new chemical reactions. Employing state of the art quantum chemical reactivity modeling methods combined with a scalable high-performance reaction database, we began to design an artificially intelligent software that will store experimental reactivity data and expand its knowledge base using computer simulations.

Accomplishments: Statement of the problem studied

The goal of this program was to create a cyber-expert software PREDICTOR capable of interacting with a synthetic organic chemist to expedite the discovery of new chemical reactions. Employing state of the art quantum chemical reactivity modeling methods combined with a scalable high-performance reaction database, we began to design an artificially intelligent software that will store experimental reactivity data and expand its knowledge base using computer simulations.

Summary of the most important results

1. We generated a collection of computational data on two series of Diels-Alder reactions. This data was intended as the core data to be used in developing the software. Ten levels of theory (different quantum chemical models) were used to acquire this data, in pursuit of a fast but reliable method. The M06-2X functional appears to give useful predictions of selectivity.
2. The Diels-Alder reactions in question were studied experimentally and appropriate solvent and temperature conditions were determined. These conditions were then to be used to collect kinetic and selectivity data for a range of reactions. Existing data in the literature for related systems was obtained with a variety of experimental conditions, making direct comparisons problematic.
3. A flexible computer infrastructure for PREDICTOR was built, residing on a physical server machine at Indiana University. A customized KNIME server allows for easy interchange of computational modules and deployment in an easy-to-use web interface. All important capabilities were implemented in prototype form or the capability to do so was demonstrated.

Training Opportunities: Nothing to Report

Results Dissemination: Nothing to Report

RPPR Final Report
as of 24-Nov-2017

Honors and Awards: Nothing to Report

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: Faculty

Participant: Dean Tantillo

Person Months Worked:

Funding Support:

Project Contribution:

International Collaboration:

International Travel:

National Academy Member:

Other Collaborators:

Participant Type: Faculty

Participant: Uttam Tambar

Person Months Worked:

Funding Support:

Project Contribution:

International Collaboration:

International Travel:

National Academy Member:

Other Collaborators:

Participant Type: Graduate Student (research assistant)

Participant: Aaron Nash

Person Months Worked:

Funding Support:

Project Contribution:

International Collaboration:

International Travel:

National Academy Member:

Other Collaborators:

Participant Type: Graduate Student (research assistant)

Participant: Christina McCulley

Person Months Worked:

Funding Support:

Project Contribution:

International Collaboration:

International Travel:

National Academy Member:

Other Collaborators:

Participant Type: Graduate Student (research assistant)

Participant: Quynh Nguyen

Person Months Worked:

Funding Support:

Project Contribution:

International Collaboration:

International Travel:

National Academy Member:

Other Collaborators:

Participant Type: Graduate Student (research assistant)

RPPR Final Report
as of 24-Nov-2017

Participant: Ryan Pemberton

Person Months Worked:

Funding Support:

Project Contribution:

International Collaboration:

International Travel:

National Academy Member:

Other Collaborators:

Participant Type: Graduate Student (research assistant)

Participant: Terrence O'Brien

Person Months Worked:

Funding Support:

Project Contribution:

International Collaboration:

International Travel:

National Academy Member:

Other Collaborators:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)

Participant: Young Hong

Person Months Worked:

Funding Support:

Project Contribution:

International Collaboration:

International Travel:

National Academy Member:

Other Collaborators:

Participant Type: Faculty

Participant: David Wild

Person Months Worked:

Funding Support:

Project Contribution:

International Collaboration:

International Travel:

National Academy Member:

Other Collaborators:

Participant Type: Graduate Student (research assistant)

Participant: Abhik Seal

Person Months Worked:

Funding Support:

Project Contribution:

International Collaboration:

International Travel:

National Academy Member:

Other Collaborators:

RPPR Final Report
as of 24-Nov-2017

Uttam K. Tambar, co-PI

Department of Biochemistry, UT Southwestern Medical Center

Introduction

Tambar's synthetic organic chemistry group worked closely with the other groups to validate the predictions made by the PREDICTOR platform through a series of laboratory experiments. For the studies related to the Diels-Alder Reaction, representative dienes and dienophiles were synthesized, and they were coupled to determine the major observed product for the various [4+2] cycloadditions.

- ❖ Organic chemists are tasked with converting simple molecules into more complex and valuable molecules, but are limited by:
 1. The extensive time it takes to design and optimize a synthesis
 2. Multi-step syntheses result in low yields of product
 3. Even lower yields of product with a desired stereochemistry are actually recovered in the end
- ❖ Computational chemistry has improved our understanding of aromaticity, reactivity, and transition states, but our ability to reliably predict stereoselectivity and designing effective synthetic routes to valuable molecules limits the rate of discovery and production.
- ❖ When considering synthetically powerful, regioselective, and stereoselective

reactions few are as impactful as the Diels-Alder reaction.

- ❖ Molecular orbital calculations support that stereoselectivity can be predicted using the “endo rule,” which assumes the transition state is stabilized by intermolecular forces caused by primary (red) and secondary (blue) π -orbital overlap, as shown in Figure 1.

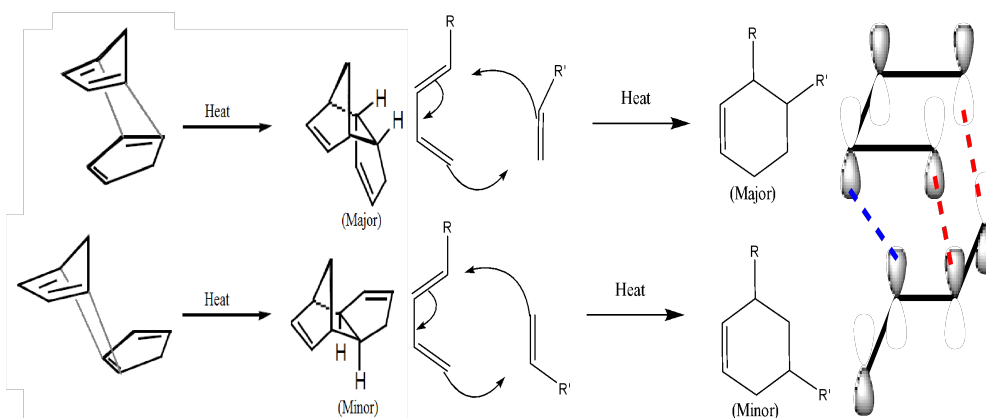


Figure 1. Diels-Alder reactions involve the cycloaddition between a diene and a dienophile via a concerted pericyclic transition state.

- ❖ The endo rule was reported as having minor importance in predicting stereoselectivity in a systematic study on reactions between cyclopentadiene and ethylene with varying methyl derivatives,⁵ which suggests that deviations from the endo rule with synthetically relevant reactants could provide access to other valuable products.
- ❖ We have developed a systematic study of Diels-Alder reactions between butadienes with varying synthetically relevant substituents and a dienophile to better determine if stereochemically powerful reactions could deviate from theoretical predictions (Figure 2).

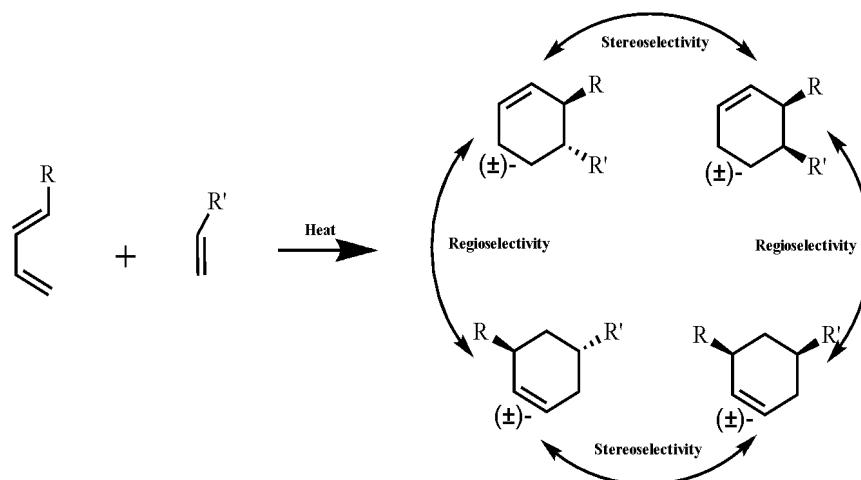


Figure 2. One reaction can form up to eight products based on the relative orientation of the diene and dienophile and on the electronic and steric properties of substituents.¹

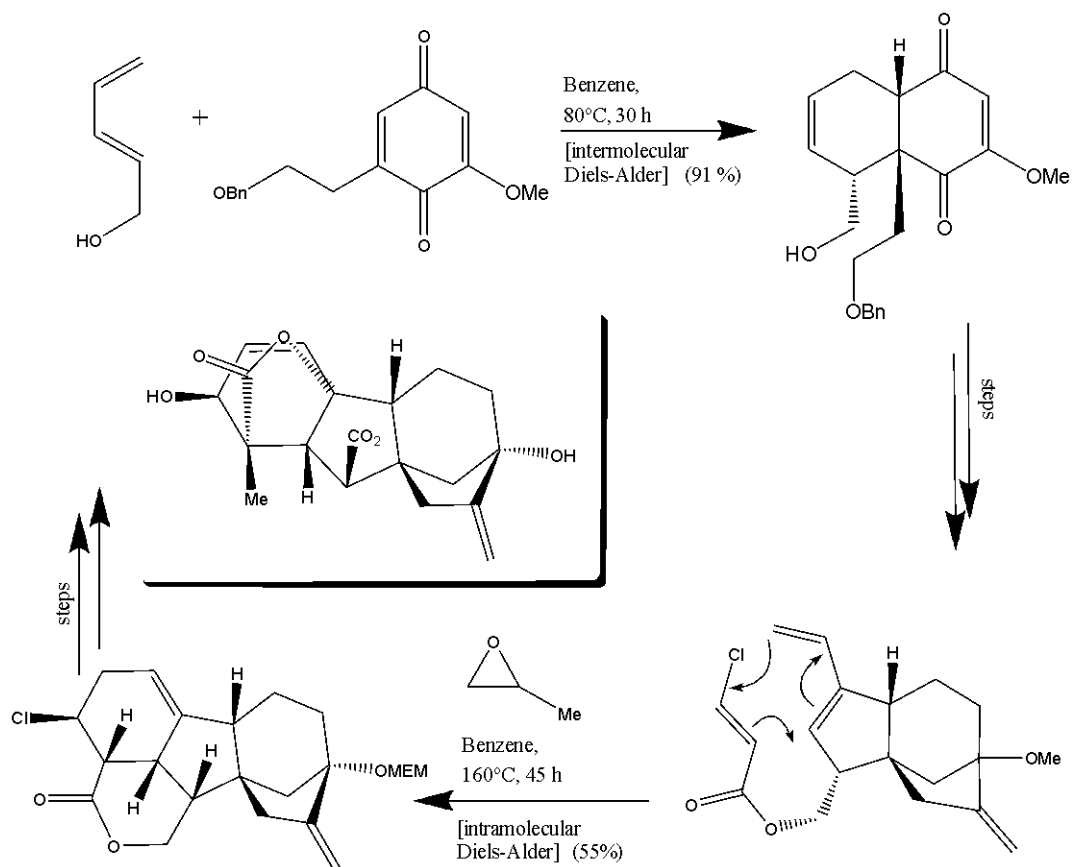


Figure 3. Synthesis of Gibberellic acid by Corey et al. (1978),² which is an important plant

growth regulator that is used commercially to increase plant growth and crop yields.

Synthesis

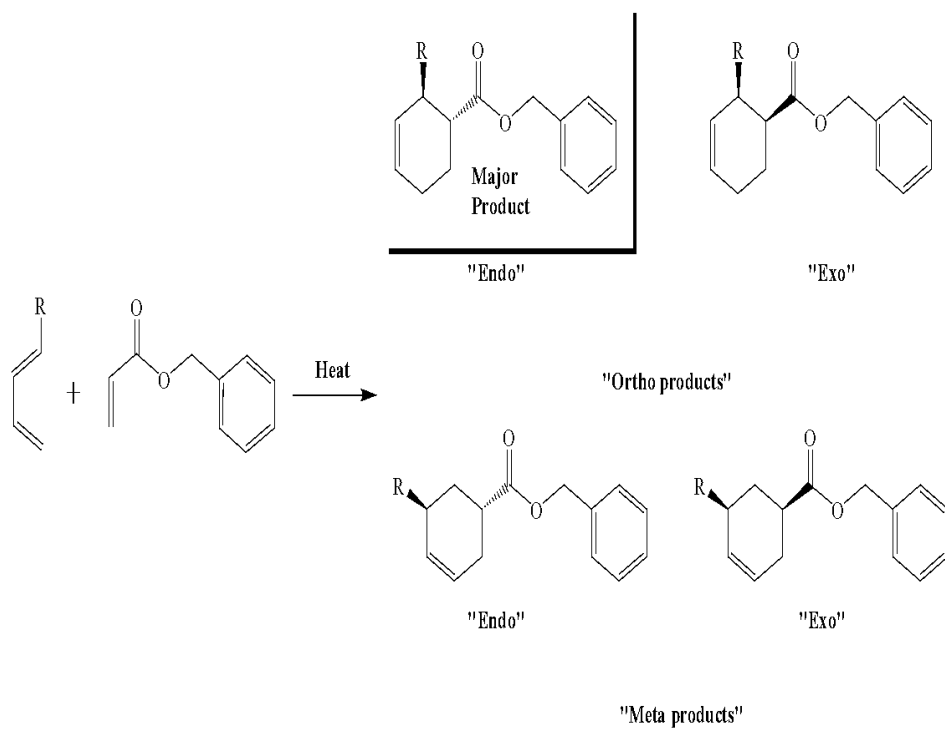


Figure 4. These substituents were used to analyze the regioselectivity and stereoselectivity of this reaction. Electron donating ability decreases from substituents 1-5.

Experimental Methods

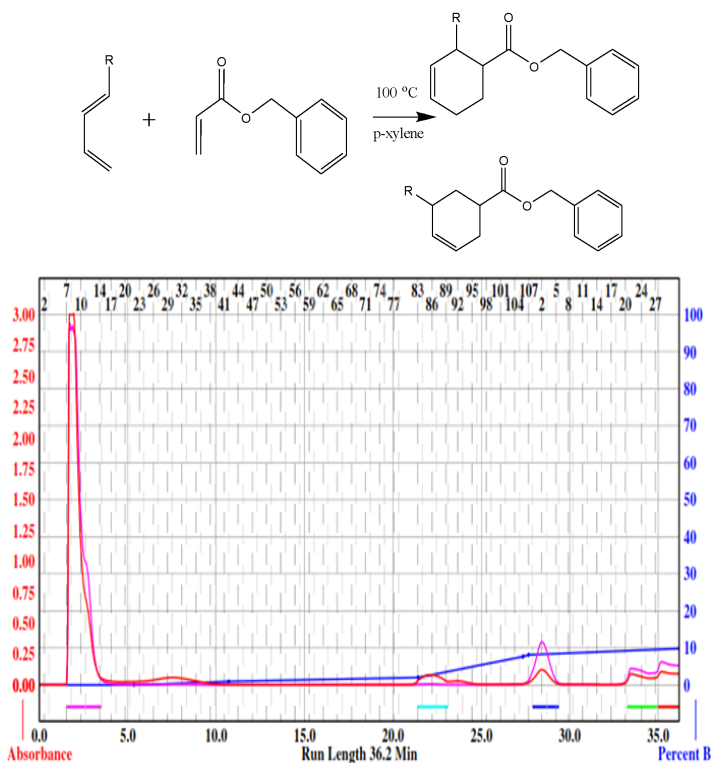
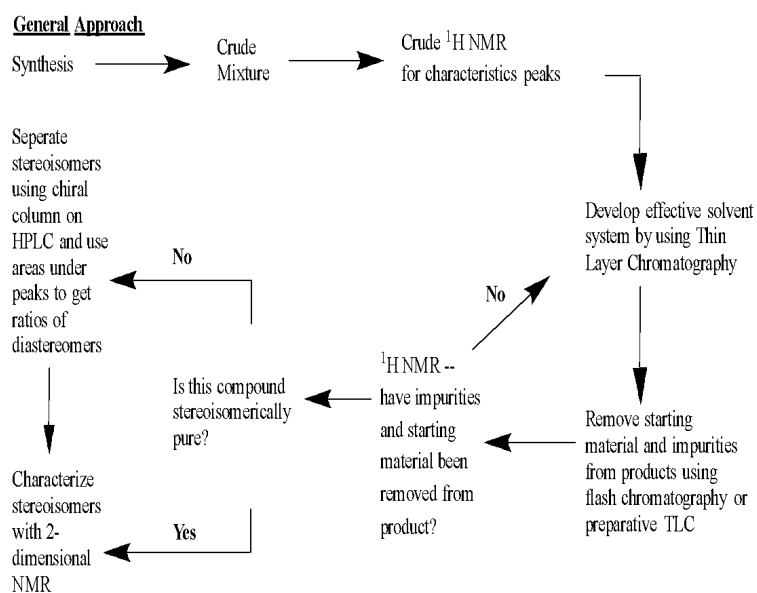
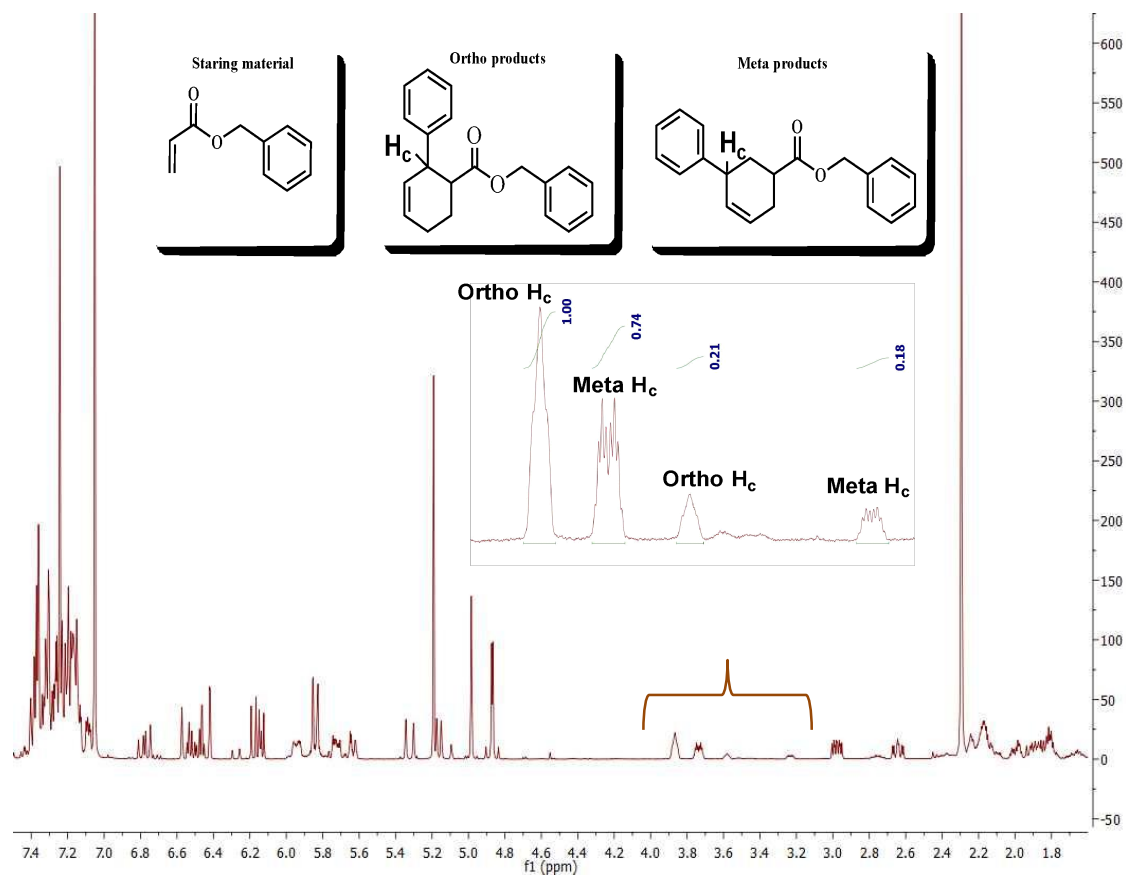


Figure 5. The presence of a benzyl ester group, which is UV active, on the dienophile allows for the easy detection and separation of products by both silica gel flash chromatography and

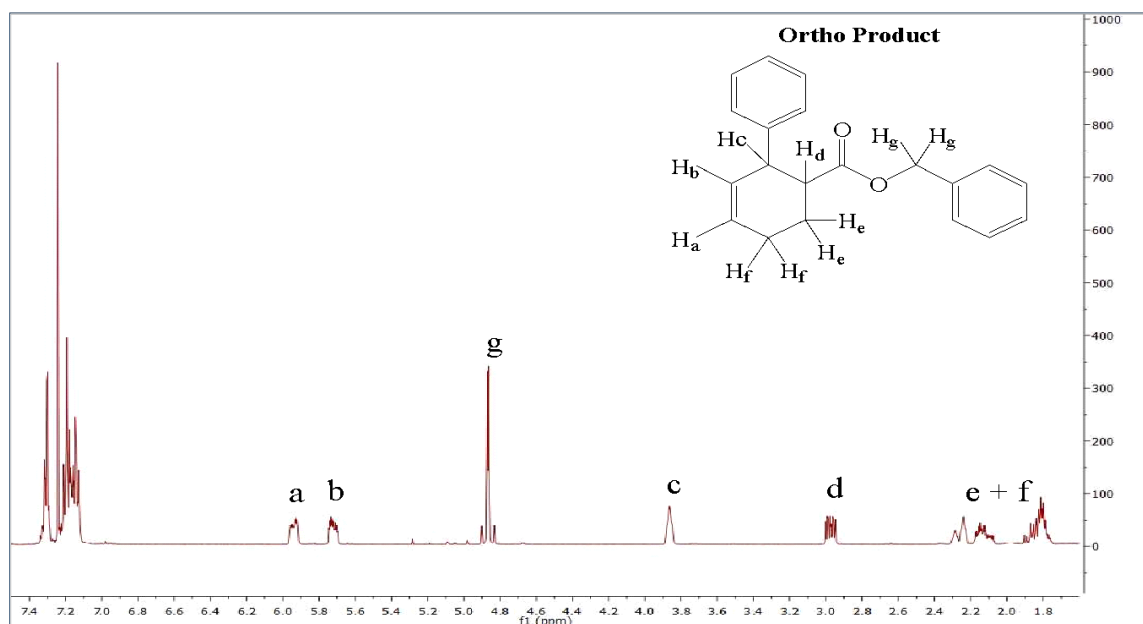
preparative HPLC.

Analysis

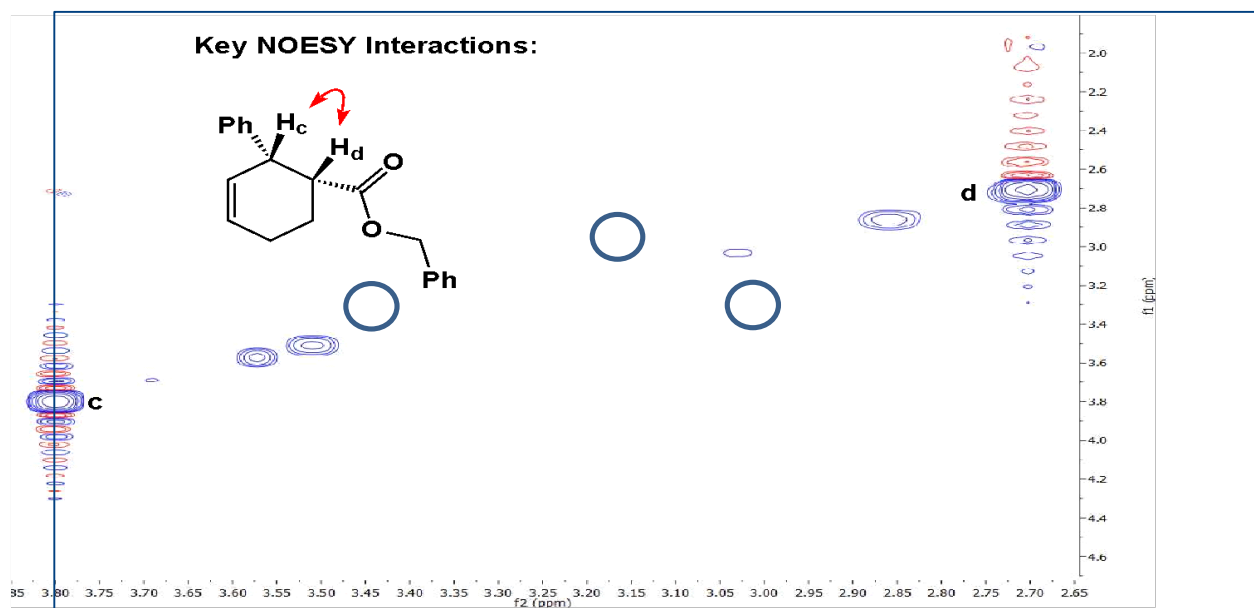
Crude ^1H NMR



^1H NMR of Major Product



2-D NMR Data of Major Product



Conclusions

- ❖ According to molecular orbital theory the primary and secondary orbital overlap provided by the benzyl ester group on the dienophile should significantly favor the

endo product for each regioisomer. Frontier molecular orbital analysis of electron withdrawing groups would also support that the regioselectivity would also strongly favor the ortho product over the meta product.

- ❖ Free energy calculations and well known textbook examples would predict that this reaction would favor the ortho over the meta product in ratios as high as 366 : 1 or in some cases even “only ortho” and “only endo.”
- ❖ The reaction between the phenyl substituted butadiene and the dineophile resulted in a regioselectivity of 5.56 : 1 favoring the ortho product. The ortho product was produced with a stereoselectivity of 4.76 : 1 favoring the endo product over the exo product. The meta product was produced with a stereoselectivity of 1.17 : 1 favoring the endo product over the exo product.
- ❖ These results suggest that while the “endo rule” has illuminated how π -orbitals govern the stereoselectivity of Diels-Alder reactions, the “endo rule” alone does not provide an accurate way to predict the outcome between substituted dienes and dienophiles.
- ❖ Future work will focus on finishing this systematic study and use these results, along with results from future systematic studies with other dienes and dienophiles, to test the accuracy of computational predictions made by our collaborators.
- ❖ The predictions will be made using PREDICTOR, a novel cyber-expert software that employs quantum chemical reactivity modeling methods and a scalable high-performance reaction database. Our short term goal is to expand the software’s knowledge with computer simulations and experimental results from systematic studies so that in the future the regioselectivity and stereoselectivity of any Diels-

Alder reaction be accurately predicted.

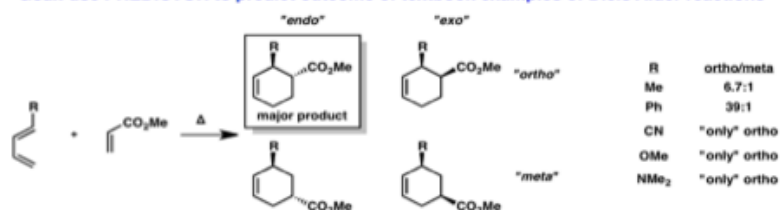
Literature cited

1. *Tetrahedron Lett.* **1982**, 23, 1139 - 1142
2. *J. Am. Chem. Soc.* **1978**, 100, 8034 - 8036.
3. <http://donohoe.chem.ox.ac.uk/resources/06042013LiteratureDi.pdf>
4. *J. Am. Chem. Soc.* **1989**, 111, 9172 – 9176.
5. *J. Am. Chem. Soc.* **1970**, 92:22, 6548 - 6553

Dean J. Tantillo, co-PI
Department of Chemistry, UC Davis

To determine the most appropriate level of theory for use in PREDICTOR, a variety of theoretical methods were applied to a series of Diels-Alder cycloadditions (Figure below). The goal was to find a level of theory that is accurate and fast. We set out to reproduce both the regioselectivity of the Diels-Alder reactions (ortho/meta ratios; first table below) and its diastereoselectivity (endo/exo ratios; second table below). We surveyed a variety of density functional theory (DFT) methods, as well as a coupled cluster method. We modeled the reactions both in the gas phase and in solvent (using the SMD continuum model). As shown below, none of the methods used provided a perfect match with experimental results. On balance, we would recommend the B3LYP/6-311++G(2d,p) level of theory, but additional theoretical methods should be surveyed.

Goal: use PREDICTOR to predict outcome of textbook examples of Diels Alder reactions



Ortho/Meta Ratios	Me	Ph	CN	OMe	NMe ₂
Experimental Results	6.7 : 1	39 : 1	only ortho	only ortho	only ortho
B3LYP/6-31G(d)	41 : 1	316 : 1	5 : 1	117 : 1	5.0x10 ⁷ : 1
M06-2X/6-311++G(2d,p)	12 : 1	5 : 1	0.2 : 1	65 : 1	1.1x10 ⁵ : 1
B3LYP/6-311++G(2d,p)	34 : 1	170 : 1	2.4 : 1	58 : 1	8.4x10 ⁷ : 1
SMD(MeCN)-B3LYP/6-311++G(2d,p)// B3LYP/6-311++G(2d,p)	40 : 1	366 : 1	28 : 1	979 : 1	2.1x10 ⁸ : 1
SMD(DCM)-B3LYP/6-311++G(2d,p)// B3LYP/6-311++G(2d,p)	47 : 1	237 : 1	16 : 1	976 : 1	4.1x10 ⁸ : 1
SMD(MeCN)-M06-2X/6-311++G(2d,p)// M06-2X/6-311++G(2d,p)	13 : 1	30 : 1	5 : 1	296 : 1	7.0x10 ⁸ : 1
SMD(DCM)-M06-2X/6-311++G(2d,p)// M06-2X/6-311++G(2d,p)	15 : 1	18 : 1	3 : 1	281 : 1	4.8x10 ⁸ : 1
ωB97-XD/6-311++G(2d,p)	37 : 1	32 : 1	0.6 : 1	223 : 1	3.1x10 ⁷ : 1
PBE1PBE/6-311++G(2d,p)	27 : 1	25 : 1	1.2 : 1	35 : 1	3.0x10 ⁶ : 1
CCSD/6-31+G(d)//B3LYP/6-311++G(2d,p)	7.3 : 1	0.3 : 1	0.006 : 1	38 : 1	3.9x10 ⁷ : 1

Endo/Exo Ratios	Me	Ph	CN	OMe	NMe ₂
B3LYP/6-31G(d)	0.4 : 1	0.1 : 1	0.2 : 1	0.6 : 1	0.4 : 1
M06-2X/6-311++G(2d,p)	0.9 : 1	0.2 : 1	0.8 : 1	5.8 : 1	26.7 : 1
B3LYP/6-311++G(2d,p)	0.5 : 1	0.1 : 1	0.3 : 1	0.5 : 1	0.5 : 1
SMD(MeCN)-B3LYP/6-311++G(2d,p)// B3LYP/6-311++G(2d,p)	1.9 : 1	2.0 : 1	3.6 : 1	1.0 : 1	0.5 : 1
SMD(DCM)-B3LYP/6-311++G(2d,p)// B3LYP/6-311++G(2d,p)	1.5 : 1	1.4 : 1	2.4 : 1	0.9 : 1	0.5 : 1
SMD(MeCN)-M06-2X/6-311++G(2d,p)// M06-2X/6-311++G(2d,p)	5.7 : 1	4.5 : 1	10 : 1	17.6 : 1	9.5 : 1
SMD(DCM)-M06-2X/6-311++G(2d,p)// M06-2X/6-311++G(2d,p)	4.5 : 1	3.4 : 1	6.4 : 1	14.2 : 1	9.7 : 1
ωB97-XD/6-311++G(2d,p)	1.9 : 1	4.1 : 1	0.9 : 1	4.7 : 1	1.9 : 1
PBE1PBE/6-311++G(2d,p)	0.5 : 1	0.2 : 1	0.3 : 1	1.2 : 1	0.8 : 1
CCSD/6-31+G(d)//B3LYP/6-311++G(2d,p)	0.3 : 1	0.9 : 1	0.4 : 1	1.7 : 1	4.8 : 1

* Values in red correspond to the level of theory with best agreement for ortho/meta ratios

Mu-Hyun Baik, co-PI

Department of Chemistry, Indiana University Bloomington

A key element of PREDICTOR is the automated acquisition and query-enabled storage of computer simulation data in a relational database. Based on our prior experience with a legacy database system called Varuna, which we developed some years ago, we have developed a new database avoiding some of the shortcomings of the legacy system. The design features of the new database were:

(1) Scalability - anticipating that a realistic size of the PREDICTOR database will quickly exceed a few million datasets, it must be designed to scale to this size without technical problems. (2) Automation - both depositing and curating the dataset entry into the database must be done automatically or in such a way requiring minimum human interaction.

Scalability and Expandability. To enable scalability, platform independence and make PREDICTOR compatible to various chemistry databases and computational tools, we decided on JAVA and SQLite as the platform of the package. SQLite has the advantage of being a scalable, SQL-compatible database system, which does not require a DB-server and, thus, can be run as a local database system. As it is fully compatible to the industry standard SQL, designing a server-based system at a later time will be trivial. Producing a JAVA front-end to the database and developing the required functions for managing computational molecular modeling jobs, automatically registering them in the database and automatically connecting and managing simulations jobs has proved to be tedious and laborious. But, during our short funding period, we were able to produce a prototype software with fully functional key components.

The main splash-display, from which all functions can be access is shown below:

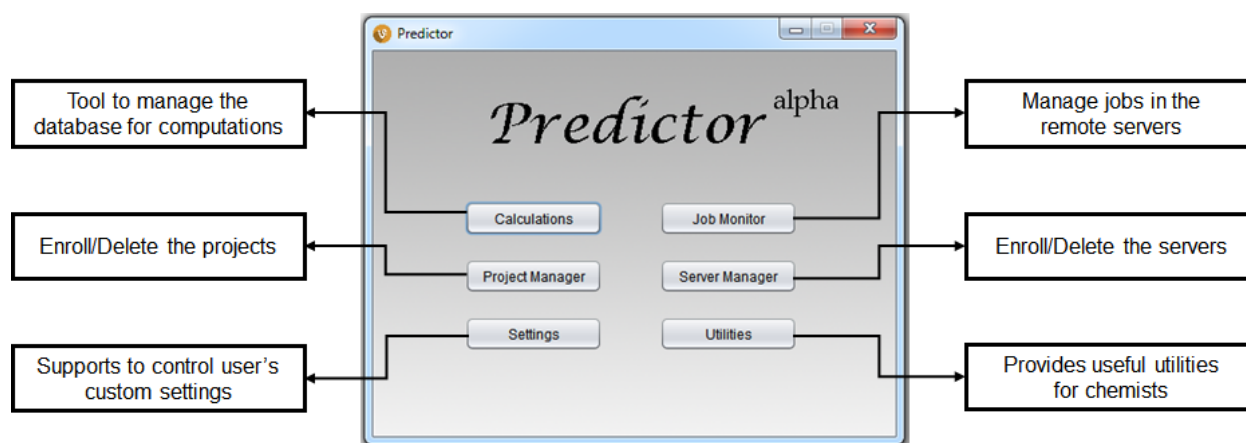


Figure 1. The general features of the PREDICTOR program.

Managing calculations and increasing scalability. The most important feature in PREDICTOR is the ability to create input files, run computer simulations on the supercomputer and process output files when simulations complete with little to no human interactions. In addition, computer simulations must be organized, data extracted and entered into the SQL-database automatically. Utilizing our legacy system, we were able to clearly identify which metadata of the calculations and which results need to be stored in the database. Key to maintaining high levels of operational efficiency while keeping enough data in the database that allows for physically meaningful queries is finding a balance between the amount of details in the metadata and computed results stored in the database vs. writing data-harvesting modules that will acquire the needed data on the fly from machine-reading the output file. We have taken advantage of the experience that we were able to gather using our legacy system VARUNA. From that experience, we know that the underlying database schema deserves some attention. After some experimentation, we have implemented the first set of core data structures, as

illustrated below.

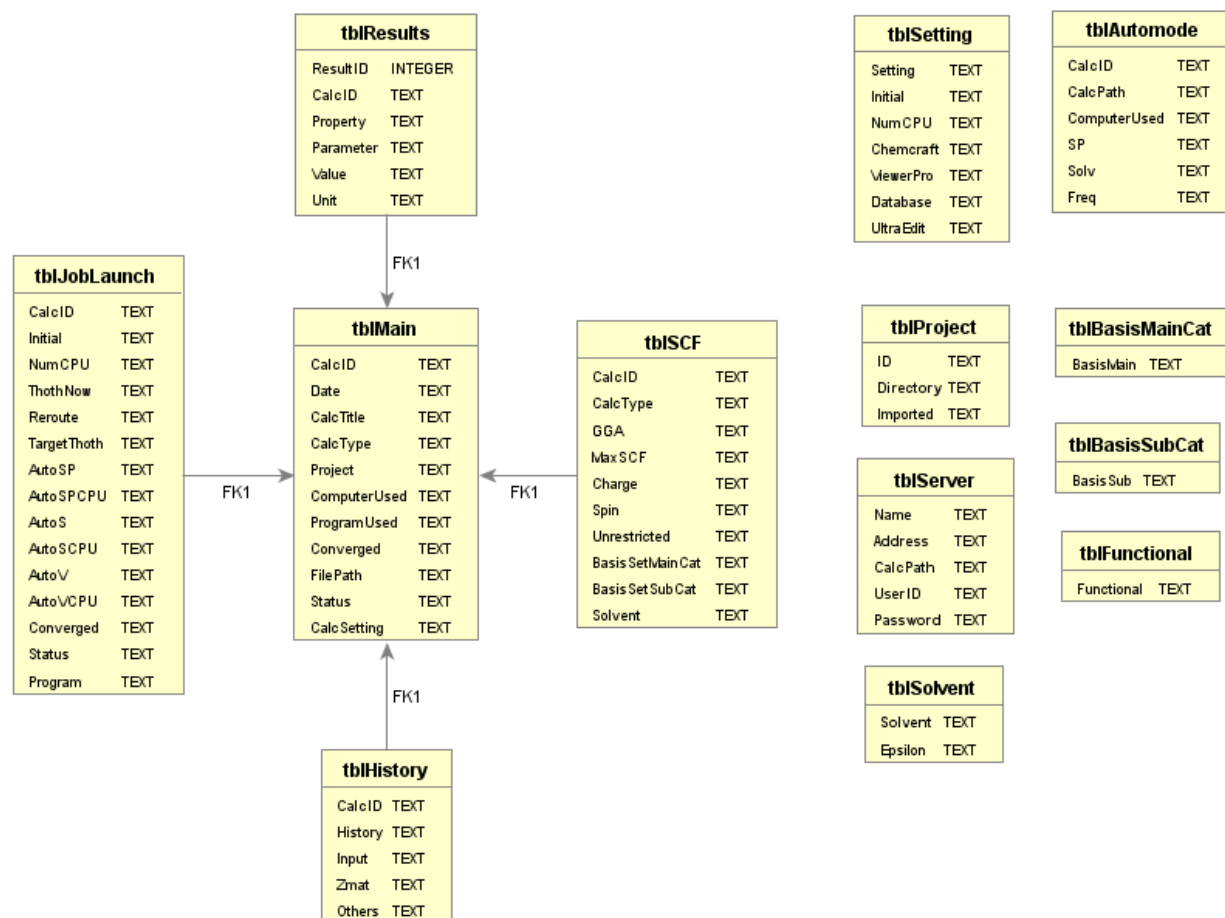


Figure 2. Database schema diagram of the external database of PREDICTOR.

Ideally, the set of the metadata of each calculation should uniquely identify the calculations. In addition to technical parameters, such as the Density Functional, basis set, simulation software used for the calculations, there are also project-based metadata, such as project for which the calculation was carried out, the file path where the output files can be found, etc. Of particular interest is the datatable **tblResults**, where key results, such as HOMO/LUMO-energies, total energies, etc. can be stored with a qualifier tag. This general storage solution allows for efficiently storing a variety of results that can be retrieved by a standard SQL-query.

Each calculation carries a Calculation-ID, which is used as a key to encode relations between the different tables of the relational database. The referential integrity of the data is maintained by the SQLite software in a professional way and the composite data is displayed in an unified frontend. We designed an effective and highly responsive graphical user interface that allows for quickly browsing through the data, as shown below.

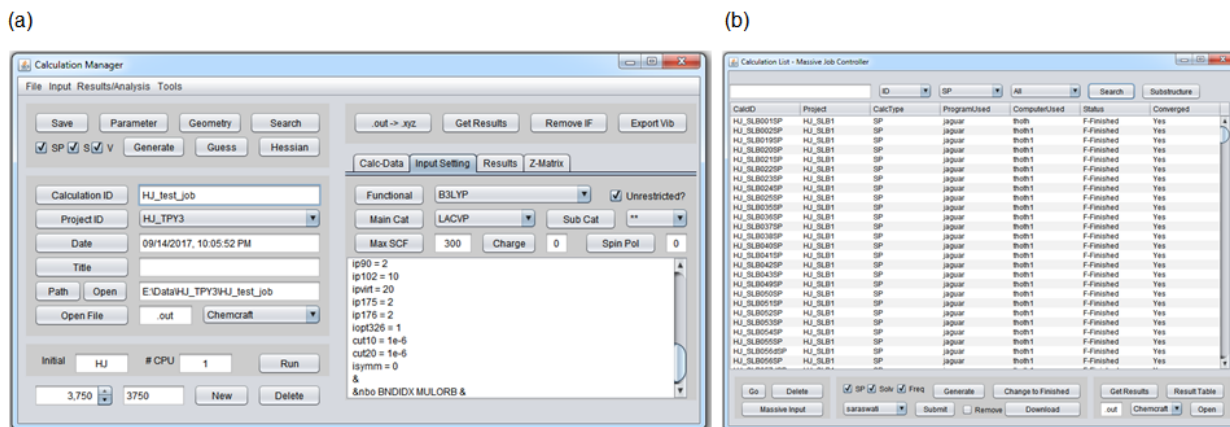
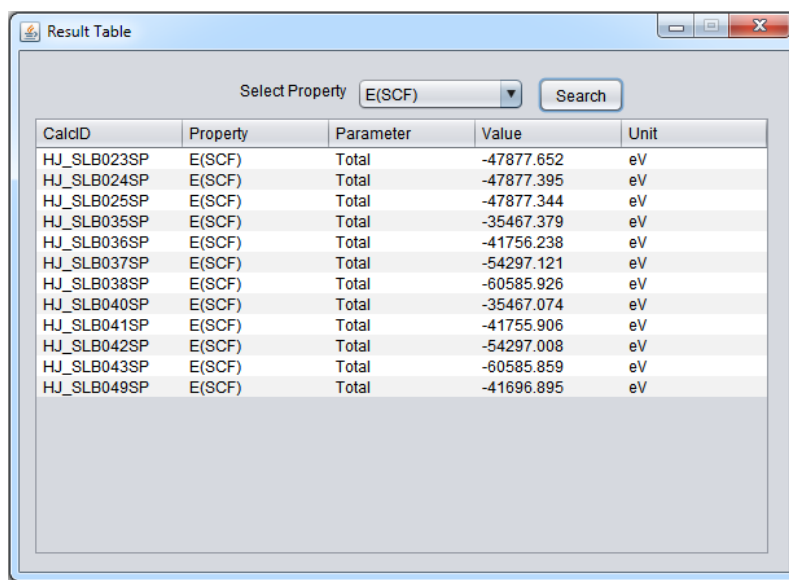


Figure 3. (a) A screenshot image of the “Calculation Manager” of PREDICTOR. (b) A screenshot image of the “Massive job controller”.

Once users creates a new calculation or a new series of calculations, the PREDICTOR retrieves all technical parameters from previous calculations, unless specified otherwise by the user. Input files for different quantum chemical engines, such as Jaguar, QChem, GAMESS, MolPro, ADF, MolCAS are automatically generated and all calculations are registered in the database. We have invested some time to program a generic SSH/SFTP-interface that allows PREDICTOR to connect to supercomputers, file-transfer, send UNIX instructions and receive the responses from the supercomputer directly. This allows for automatically launching and monitoring the progress of computer simulations. Although some additional features such as automatic error corrections remains to be further improved, much of the “Auto-Mode”

functionality has been completed and PREDICTOR can already carry out entire series of calculations consisting of several thousands of calculations without human interaction. A list of calculations that are being monitored is displayed and manual intervention is conveniently possible, as shown in a screen capture in Figure 3b.



CalcID	Property	Parameter	Value	Unit
HJ_SLB023SP	E(SCF)	Total	-47877.652	eV
HJ_SLB024SP	E(SCF)	Total	-47877.395	eV
HJ_SLB025SP	E(SCF)	Total	-47877.344	eV
HJ_SLB035SP	E(SCF)	Total	-35467.379	eV
HJ_SLB036SP	E(SCF)	Total	-41756.238	eV
HJ_SLB037SP	E(SCF)	Total	-54297.121	eV
HJ_SLB038SP	E(SCF)	Total	-60585.926	eV
HJ_SLB040SP	E(SCF)	Total	-35467.074	eV
HJ_SLB041SP	E(SCF)	Total	-41755.906	eV
HJ_SLB042SP	E(SCF)	Total	-54297.008	eV
HJ_SLB043SP	E(SCF)	Total	-60585.859	eV
HJ_SLB049SP	E(SCF)	Total	-41696.895	eV

Figure 4. A captured image of the result table - a massive calculation result management tool.

We have implemented several key functionalities in data retrieval and analysis that were previously found to be essential in our modeling studies. The data retrieval model of PREDICTOR distinguishes two different facilities: (a) A few selected key indicators, such as HOMO/LUMO and electronic energies are imported and stored in the database and can be searched using SQL-queries (b) More complex data points, such as partial charges, the coefficients of molecular orbitals, bond distances in the molecules, are not stored in the database, but are retrieved when needed by machine-reading the output files and gathering the data on demand. As the location of the output files are stored in the database, this can be done automatically. As the volume of the data grows, the ad hoc gathering of new data points will

require some time and it will therefore be important to make this process as efficient as possible. PREDICTOR can already read thousands of lines in an output file quickly and gather valuable information much more efficiently than any manual extraction protocol could. Figure 4 shows the result browsing capabilities - of course, the result browsing and exploration modules can be developed to a much higher level of sophistication. During this funding period, we aimed to establish the foundation first, upon which we will build many more sophisticated analysis tools in the future.

Job control and automation. An important feature in PREDICTOR is that users can submit the simulations jobs directly to a server/supercomputer without having to manually transfer input files or manipulated job submission scripts. These trivial tasks typically consume a significant amount of time during each work day of a computational modeller and it is straightforward to implement an automatic workflow. PREDICTOR transfers all job files to the server, prepares the necessary job launching scripts, runs and monitors jobs with little to no human supervision. Instead of launching one job at a time using the UNIX terminal, PREDICTOR allows for one user to launch and monitor thousands of jobs automatically. To enable maximum flexibility and new features in the future, we implemented a generic SSH and SFTP communication channel, where PREDICTOR can communicate with supercomputers directly.

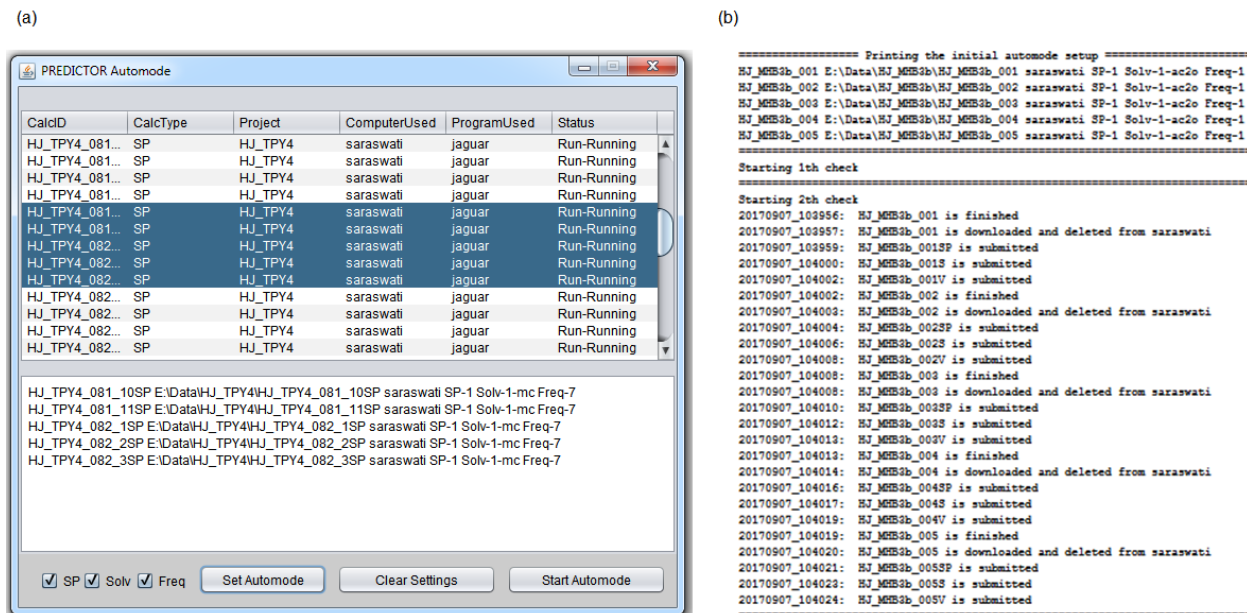


Figure 5. (a) A captured image of the result table - a massive calculation result management tool. (b) An image of the Auto-Mode logfile.

One important feature that we implemented already is “Auto-Mode”. When running a series of calculations, the need often arises where follow-up calculations are necessary. For example, successful geometry optimizations may required vibrational frequency calculations to be run, or solvation calculations to be carried out. And there are also a handful of “standard cases” where a calculation may have failed for a well-known reason. For PREDICTOR to become a virtual expert, it must know how to react to these cases. And, thus, we have developed an “Auto-Mode”, where PREDICTOR will carry out these follow-up calculations automatically. This tool is tremendously helpful in its preliminary implementation already, as the user no longer has to monitor and carry out these routine calculations manually. To keep the operational model flexible, our current implementation allows the user to enable Auto-Mode for specific calculations. That is, the user can currently choose which calculations will be monitored and handled by PREDICTOR and which are not monitored. To the best of our knowledge,

PREDICTOR is currently the only software capable of such a flexible and automated job handling.

Substructure Searching. Although the implementation of the features mentioned above and the design of the new database required significant effort, these features were available in the legacy software VARUNA in some form and we had a clear vision on how to implement these features. An entirely new and challenging feature that we desired is the capability of recognizing and searching for substructures. For the envisioned development of an artificially intelligent program that can make chemically meaningful decisions based on quantum chemical simulations, it is of the utmost importance that PREDICTOR is aware of chemical structures and functional groups. While the 3-dimensional structure is easily ascertained by analyzing the Cartesian coordinates of the molecules, we need to be able to group molecules together based on chemically meaningful similarities in structure and composition. And this must be done in a computationally efficient manner.

Hence, we implemented a substructure searching functionality based on a structural fingerprint of each molecule that is generated against a dictionary of substructures. In short, each molecule is examined at the time when it is deposited into the database and a bit-string is generated where the presence or absence of common functional groups is encoded. This is a well-known method for structure-encoding and is used for example in the Cambridge Crystal Structure Database for substructure searches. We have adopted the same protocol for our purposes based on an open source libraries provided in the CDK (Chemistry Development Kit) and Openbabel. There is still much left to do and we will need to experiment with different algorithms to identify the best way of recognizing and manipulating molecular structures, but the initial substructure searching capabilities have been implemented. Figure 6 summarized one

simple substructure search that we have implemented using the above mentioned libraries. We first derive molecular fingerprints based on SMARTS. The challenge lies in converting XYZ coordinates to SMARTS - for simple structures, Openbabel is a good starting point. As our quantum chemical database has very specific needs, such as deriving a similarity index based on compatible electronic structures and computed chemical properties, we are likely going to expand these algorithms - work in this direction is planned.

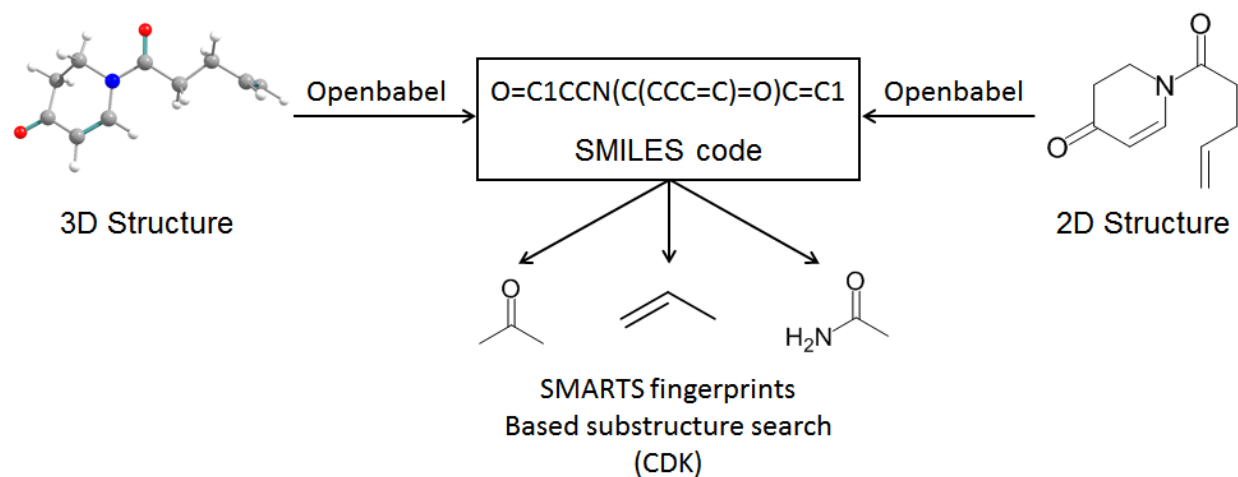


Figure 6. Strategies for the substructure search using external libraries.